A Fundamental Framework for Realizing Semantic Interoperability in a Multidatabase Environment

Yasushi Kiyoki, Takashi Kitagawa and Youichi Hitomi

Institute of Information Sciences and Electronics
University of Tsukuba
Tsukuba 305 Japan

phone: 81-298-53-5187, fax: 81-298-53-5206
e-mail: kiyoki@is.tsukuba.ac.jp, takashi@is.tsukuba.ac.jp

Abstract

In multidatabase research, the realization of semantic interoperability is the most important issue for resolving semantic heterogeneity between different databases.

In this paper, we propose a fundamental framework for realizing semantic interoperability at the level of semantic relationships between data items in a multidatabase environment. We present a metadatabase system which extracts the significant information from different databases. The metadatabase system uses a mathematical model of meaning to dynamically recognize the semantic equivalence, similarity and difference between data items. The essential feature of this model is that the specific meaning of a data item is dynamically fixed and unambiguously recognized according to the context by semantic interpretation mechanisms.

Key Words: multidatabases, semantic interoperability, metadatabases, semantic equivalence, semantic interpretation

1 Introduction

In recent years, a number of databases have become available through world-wide computer networks and database users now have environments for extracting the shared information from those databases. However, it is often difficult to select the appropriate databases and extract the significant information from those databases.

Multidatabases and interoperable systems are important areas of current research concerning information resource sharing[1, 2, 8, 17]. Multidatabases are classified into four types: global schema multidatabases, federated databases, multidatabase language systems and homogeneous multidatabase language systems[2].

The current issues in multidatabase research have been summerized in [2, 17]. Basically, they are classified into "site autonomy," "differences in data representation," "heterogeneous local databases," "global constraint," "global query processing," "concurrency control," and "security." In particular, "site autonomy" and "differences in data representation" are the essential problems in implementing multidatabase environments.

This paper deals with the problem of the differences in data representation and meaning in multidatabase environments. The differences in data representation and meaning include the concrete problem of "name differences." As there are many ways to represent a given real-world object and its relationships to other objects, the same object might be represented in different ways in different databases. In general, it is

often difficult to deal with such differences in multidatabase environments. The main reason for this difficulty is the lack of knowledge of each database. Database users may not know the contents of the databases sufficiently (the fields of the databases, the available data types, the meaning of each keyword, and so on). Therefore, database users may not be able to use appropriate information in constructing their questions.

The basic problem in conventional database systems is that the fundamental operation for extracting information is string-based pattern-matching between data items. This pattern-matching cannot be effectively used to find semantically equivalent data items with different data representations, and to recognize the different meanings inherent in a data item. Furthermore, in conventional database systems, in order to provide the facilities for sharing information, the framework for the abstraction of information and the structuring of information is prescribed by the schema design. Since the structuring and abstraction of information are statically fixed at the schema design, it is difficult to access databases without knowledge of the structuring and abstraction. Utilizations of conventional database systems are prescribed by static structuring, abstraction and the string-based pattern matching. Several approaches for dealing with extended pattern matching have been proposed[14, 4].

In general, dealing with the meanings of data items in a multidatabase environment is more complicated than in that of a single database. One of the hardest problems is that it is difficult to identify the semantic equivalence, similarity and difference between data items which are extracted from different databases [2, 7, 18]. It is not easy for

users to select the appropriate databases and extract significant information for their requests. To provide the facilities for selecting the appropriate databases and extracting the significant information from those databases, a methodology for realizing semantic interoperability is an important part of database integration technology [3, 8, 10, 16, 18, 19, 20].

The problematic relationships between data items for realizing semantic interoperability are classified into two types: "homonyms" and "synonyms." Homonym means the same data item is used for different concepts. Synonym means the same concept is described by different data items in different databases.

In this paper, we propose a fundamental framework for realizing semantic interoperability in a multidatabase environment. We present a metadatabase system which extracts the significant information from different databases. This system dynamically recognizes the semantic equivalence, similarity and difference between data items which are included in different databases. We have classified metadata items into three levels. The metadatabase system includes those three levels of information independently. These levels are referred to as the "overview level", the "attribute level" and the "attribute value level." In these levels, mappings between data items in different databases are essentially needed to realize semantic interoperability. In this system, a mathematical model of meaning is used to realize semantic interoperability among the data items. This model is used to find semantically equivalent or similar data items with different data representations and to recognize the different meanings of a data

item. The main feature of this model is that the specific meaning of a data item can be dynamically fixed and unambiguously recognized according to the context.

We assert that as the semantic equivalence, similarity and difference are context dependent, computational mechanisms for interpreting semantics according to the given context must be provided. The approach used in this model to compute the distance between data items is quite different from the approaches using the concept of probability and the fuzzy theory. In this model, the meanings of data items and distances between those data items are fixed in a context dependent way and are dynamically computed and unambiguously recognized according to the given context. In this system, a keyword is issued with its context words which explain and fix the meaning of the keyword, and the data items with the equivalent or closest meaning to the keyword are extracted.

In this system, the computation for obtaining the data item with the equivalent or closest meaning to the keyword is performed very fast, because the data items with the closest meaning are physically located at the closest position in the dynamically selected semantic space.

2 Metadatabase System

2.1 An Overview of the Metadatabase System

We have presented the configuration of the metadatabase system in [11, 19]. The overview of the system is given below. The metadatabase system selects appropriate

databases for requests of database users by using metadata items and primitive functions. The structure of this system is shown in Figure 1. This system consists of the following subsystems:

(1) Database Selection Subsystem: This subsystem supports the facilities for selecting appropriate databases. It extracts the names of databases which are used for the given query and creates database keywords used in the selected databases and outputs them along with the original query to the Query Transformation Subsystem. This subsystem includes the module of the mathematical model of meaning which supports semantic interoperability, as shown in Figure 2.

(2)Query Transformation Subsystem: This subsystem supports the facilities for transforming a user's query into local queries for the selected databases. It issues the transformed queries to the selected databases.

(3) Result Integration Subsystem: This subsystem provides the facilities for integrating all the results which are obtained from the selected databases.

(4) Metadatabase Manipulation Subsystem: This subsystem supports the facilities for creating metadatabases and keeping them consistent.

(5) Metadata Acquisition Subsystem: This subsystem supports the facilities for acquiring metadata and updating the metadata by computing statistics for each database.

The Database Selection Subsystem finds the appropriate databases from which the system extracts results for the queries. The appropriate databases are selected by using metadata concerning each database. Our system provides a set of basic data items and

a set of basic functions to select the appropriate databases. The metadata items and basic functions are classified into three levels in order to perform the database selection systematically. These data items and functions are called "metadata items" and "basic functions."

2.2 Metadatabases in a Multidatabase Environment

Metadata items are classified into three levels as follows:

1. Overview level: Metadata items of this level represent common characteristics of all the data items in a database and are used to find appropriate databases without referring to the contents of individual databases. The overview information of each database includes the field of the stored contents, the fields of the database, statistical information of the database, producer names and the time coverage.

2. Attribute level: The attribute information includes the data items, such as attribute names, explanatory information and features of the attribute. The attribute information is used to judge whether or not the appropriate attributes exist in the databases. For this judgement, it is required to check for conflicts in the representation of the same semantics in different databases.

3. Attribute value level: The information concerning attribute values which is extracted from individual databases is included as data items. The attribute-value information is used to judge whether or not the appropriate attribute values

are included in the databases. For this judgement, it is required to recognize attributes values with the same semantics in different databases.

To find different data items with equivalent or similar meanings and to recognize the different meanings of a data item, the metadatabase system needs to deal with the meanings of those data items unambiguously. We introduce a new model to unambiguously recognize the specific meaning of a data item according to the context.

3 Mathematical Model of Meaning

3.1 Outline of the Model

In this section, the outline of the mathematical model of meaning is presented. The rigorous mathematical formulation is given in the next subsection. An actual example applied to multidatabase system is given in Section 6.

1. Assumption :

   A set of data which characterizes the words to be used is given in the form of an m by n matrix. In other words, for m given words, each word is characterized by n features. An automatic machinary to construct the matrix is presented in Section 4. We refer to the matrix data as the matrix A, hereafter.

2. Defining Image Space $\mathcal{I}$:

First we construct the correlation matrix with respect to the features. Then we execute the eigenvalue decomposition of the correlation matrix and normalize the eigenvectors. We define the image space $\mathcal{I}$ as the span of the eigenvectors which correspond to nonzero eigenvalues. We call such eigenvectors semantic elements hereafter. We note that since the correlation matrix is symmetric, the semantic elements form orthonormal bases for $\mathcal{I}$. The dimension ν of the image space $\mathcal{I}$ is identical to the rank of the data matrix A. Since $\mathcal{I}$ is the ν dimensional Euclidean space, various norms can be defined and a metric is naturally introduced.

3. Defining a Set of Semantic Projections Π_ν:

   We consider the set of all the projections from the image space $\mathcal{I}$ to the invariant subspaces (eigen spaces). We refer to the projection as the semantic projection and the corresponding projected space as the semantic subspace. Since the number of i dimensional invariant subspaces is $(\nu(\nu-1)\cdots(\nu-i+1))/i!$, the total number of the semantic projections is 2^ν. That is, this model can express 2^ν different phases of the meanings.

4. Constructing the Semantic Operator S_p:

   Suppose a sequence s_ℓ of ℓ words which determines the context is given. We construct an operator S_p to determine the semantic projection according to the context. Hereafter, we refer to the operator as a semantic interpretation operator or a semantic operator for brevity.

(a) First we map the ℓ words to the image space $\mathcal{I}$. This mathematically means that we execute the Fourier expansion of the sequence s_ℓ in $\mathcal{I}$ and seek the Fourier coefficients of the words with respect to the semantic elements. This corresponds to seeking the correlation between each word of s_ℓ and each semantic element.

(b) Then we add the values of the Fourier coefficients for each semantic element. This corresponds to finding the correlation between the sequence s_ℓ and each semantic element. Since we have ν semantic elements, we can constitute a ν dimensional vector. We call the vector normalized in the infinity norm the semantic center $\mathbf{G}^+(s_\ell)$ of the sequence s_ℓ.

(c) If the sum obtained in (b) for a semantic element is greater than a given threshold, ε_s, we employ the semantic element to form the projected semantic subspace. We define the semantic projection by the sum of such projections.

This operator automatically selects the semantic subspace which is highly correlated with the sequence s_ℓ of the ℓ words which determines the context.

This model can be implemented on the computer system and makes dynamic semantic interpretation possible. We emphasize here that, in our model, the "meaning" is the selection of the semantic subspace, namely, the selection of the semantic projection and the "interpretation" is the best approximation in the selected semantic subspace.

3.2 Formulation

In this section we develop the corresponding rigorous mathematical formulation. Refer to the previous subsection for the intuitive explanation of each formulation. The subsection numbers exactly correspond to those of the previous subsection.

3.2.1 Assumption

Suppose m words are given and each word is characterized by n features $(f_1, f_2, \cdots, f_n)$. For given $\mathbf{w}_i (i = 1, \cdots, m)$, we set the data matrix A to be the $m \times n$ matrix whose i-th row is $\mathbf{w}_i$ (Figure 3).

3.2.2 Defining the Image Space $\mathcal{I}$

1. We make the correlation matrix $A^T A$ of A, where A^T represents the transpose of A.

2. The eigenvalue decomposition of $A^T A$.

$$A^T A = Q \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_\nu & \\ & & & {}^{0.}{}_{0} \end{pmatrix} Q^T,$$

   where $0 \leq \nu \leq n$.

   The orthogonal matrix Q is defined as:

$$Q = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n)^T,$$

where the $\mathbf{q}_i$'s are the normalized eigenvectors of A^TA. We refer the eigenvectors as semantic elements hereafter. We note here that all the eigenvalues are real and all the eigenvectors are mutually orthogonal because the matrix A^TA is symmetric.

3. Defining image space $\mathcal{I}$.

$$\mathcal{I} := span(\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_\nu),$$

which is a linear space generated by the linear combinations of $\{\mathbf{q}_1, \cdots, \mathbf{q}_\nu\}$. We note that $\{\mathbf{q}_1, \cdots, \mathbf{q}_\nu\}$ are the orthonormal bases of $\mathcal{I}$.

3.2.3 The Set of Semantic Projections Π_ν

We first define the projection P_{λ_i} as the projection to the eigenspace corresponding to the eigenvalue λ_i,

i.e. $P_{\lambda_i} : \mathcal{I} \rightarrow span(\mathbf{q}_i)$.

We define the set of the semantic projections Π_ν as follows:

$$\Pi_\nu :=$$

$$\{\ 0\ ,\ P_{\lambda_1}, P_{\lambda_2}, \cdots, P_{\lambda_\nu},$$

$$P_{\lambda_1} + P_{\lambda_2}, P_{\lambda_1} + P_{\lambda_3}, \cdots, P_{\lambda_{\nu-1}} + P_{\lambda_\nu},$$

$$\vdots$$

$$P_{\lambda_1} + P_{\lambda_2} + \cdots + P_{\lambda_\nu}\}.$$

The number of the elements of Π_ν is 2^ν, and accordingly this implies that 2^ν different phases of meaning can be expressed by this formulation.

3.2.4 Construction of Semantic Operator S_p.

Given a sequence

$$s_\ell = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell)$$

of ℓ words which determines the context and a positive real number $0 < \varepsilon_s < 1$, the semantic operator S_p constitutes a semantic projection $P_\varepsilon(s_\ell)$, according to this context. That is,

$$S_p : T_\ell \longmapsto \Pi_\nu$$

where T_ℓ is the set of sequences of ℓ words and $T_\ell \ni s_\ell$, and $\Pi_\nu \ni P_\varepsilon(s_\ell)$. Note that the set $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell\}$ must be the words defined by the features in the matrix A.

The constitution of the operator S_p consists of the following processes:

1. Fourier expansion of $\mathbf{u}_i (i = 1, 2, \cdots, \ell)$.

   We set the inner product of $\mathbf{u}_i$ and $\mathbf{q}_j$ u_{ij}, i.e.

   $$u_{ij} := (\mathbf{u}_i, \mathbf{q}_j)\ ,\ for\ j = 1, 2, \cdots, \nu.$$

   We define $\widehat{\mathbf{u}}_i \in \mathcal{I}$ as

   $$\widehat{\mathbf{u}}_i := (u_{i1}, u_{i2}, \cdots, u_{i\nu}).$$

This is the mapping of the word $\mathbf{u}_i$ to the image space $\mathcal{I}$. The correlations between the word and each semantic element are computed by this process.

2. Computing the semantic center $\mathbf{G}^+(s_\ell)$ of the sequence s_ℓ.

$$\mathbf{G}^+(s_\ell) := \frac{\left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right)}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu}\right) \|_\infty},$$

where $\| \cdot \|_\infty$ denotes the infinity norm. We note that the semantic center $\mathbf{G}^+(s_\ell)$ is a ν dimensional vector and the i-th element represents the correlation between the context words s_ℓ and the i-th semantic element $\mathbf{q}_i$.

3. Determining the semantic projection $P_\varepsilon(s_\ell)$.

$$P_\varepsilon(s_\ell) := \sum_{i \in \Lambda \varepsilon_s} P_{\lambda_i} \quad \in \Pi_\nu,$$

where $\Lambda_\varepsilon := \{\ i \mid |\left(\mathbf{G}^+(s_\ell)\right)_i \mid > \varepsilon_s\}$.

Thus, we can construct the operator S_p which determines a projection $P_\varepsilon(s_\ell)$ from the context words s_ℓ.

3.2.5 The Dynamic Metric $\rho\left(\mathbf{x}, \mathbf{y}; s_\ell\right)$

We introduce a dynamic metric $\rho\left(\mathbf{x}, \mathbf{y}; s_\ell\right)$ for $\mathbf{x}, \mathbf{y} \in \mathcal{I}$, which dynamically changes depending on the context. This metric is designed in order that the model faithfully reflects the change of the context. The metric $\rho\left(\mathbf{x}, \mathbf{y}; s_\ell\right)$ is defined as follows:

$$\rho\left(\mathbf{x}, \mathbf{y}; s_\ell\right) = \sqrt{\sum_{j \in \Lambda_\varepsilon} \left\{c_j(s_\ell)\left(x_j - y_j\right)\right\}^2},$$

where the weight $c_j(s_\ell)$ is given by:

$$c_j(s_\ell) := \frac{\sum_{i=1}^{\ell} u_{ij}}{\| \left(\sum_{i=1}^{\ell} u_{i1}, \cdots, \sum_{i=1}^{\ell} u_{i\nu} \right) \|_\infty},$$

$$j \in \Lambda_\varepsilon.$$

Note that this metric significantly reduces the dependency on the selection of the bases for the image space and on the selection of the threshold ε_s for determining the corresponding semantic subspace. This is because that the weight $c_j(s_\ell)$ automatically reduces the effects of the semantic elements with small significance.

4 Creation of the Image Space

In our implementation study for the creation of the image space, we have referred to the English dictionary "General Basic English Dictionary [15]," in which only 850 basic words are used to explain every English vocabulary entry. These basic words were used as features, that is, they were used as the features corresponding to the columns of the data matrix A. Namely, 850 features were provided to make the image space. In addition, 2000 words were used to represent the words corresponding to the rows of the data matrix A. These 2000 words have been selected as the basic vocabulary entries. These entries are the basic explanatory words used in the English dictionary "Longman Dictionary of Contemporary English [9]." The 2000 $\times$ 850 data matrix was used to create the image space.

By using this matrix, an image space is computed within the framework of the

mathematical model of meaning. This space represents the semantic space for computing the meanings of the keywords and data items which are used in a multidatabase environment. A given keyword and data items in the three metadatabase levels are mapped to this space. The image space is created independently of the contents of the individual databases. Furthermore, each basic word corresponding to the vocabulary entries is mapped to the image space by the Fourier expansion. The procedure for the creation of the image space is as follows:

1. Each of the 2000 vocabulary entries corresponds to a row of the matrix A. When defining a row of the matrix A, each column corresponding to the explanatory words (features) which appear in each vocabulary entry is set to the value "1". If features of the English word are used in the negative, the columns corresponding to those features are set to the value "-1". All the other columns are set to the value "0". This process is performed for every vocabulary entry. Then, each column of the matrix is normalized by the 2-norm to create the data matrix A.

2. By using this matrix A, the image space is computed as described in Section 3. This space represents the semantic space for computing the meanings of the data items.

To create the data matrix A from the dictionary automatically, we have implemented several filters which remove unnecessary words, such as articles and pronouns, and transform conjugations and inflections of words to the infinitives. The unnecessary

words are not used as features in the data matrix A. The process of creating the data matrix A is shown in Figure 4. Each filter has the following functions:

1. filter-1: This filter inputs the original text which explains an English vocabulary entry in the dictionary and transforms the capital letters to small ones. Then, it outputs the filtered text to filter-2.

2. filter-2: This filter removes the special symbols, such as colons and semicolons. Then, it outputs the filtered text to filter-3.

3. filter-3: This filter eliminates the unnecessary words, such as articles and pronouns, from the text. Then, it outputs the filtered text to filter-4.

4. filter-4: This filter transforms conjugations and inflections to the infinitives. This filter also transforms abbreviated letters to original words. Then, it outputs the filtered text to filter-5.

5. filter-5: The rows of the data matrix A are created for each English vocabulary entry by setting the filtered words to the features.

5 Basic Functions for Semantic Interoperability

The mathematical model of meaning is realized as a module of the metadatabase system in order to support semantic interoperability at the level of the semantic relationships between data items in a multidatabase environment. This module extracts semantically

equivalent or similar data items with different data representations and recognizes the different meanings of a data item among the different databases. Three basic functions are provided to compute the meanings of the data items, as shown in Figure 5.

5.1 Function-1: Selection of the Semantic Subspace

Given the sequence

$$s_\ell = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_\ell)$$

of context words used for determining the context, its semantic center and semantic projection are computed and the selection of the semantic subspace is performed. Each word $\mathbf{u}_i$ is defined by the features in the data matrix A.

This selection determines the context defined by the given context words. This function is implemented by the following procedure:

Step-1: Fourier expansion of the context words.

The Fourier expansion is executed for each context word $\mathbf{u}_i$, and the Fourier coefficients of these context words are computed for each semantic element. This corresponds to computing the correlation between each context word and each semantic element.

Step-2: The addition of the Fourier coefficients of each individual semantic element.

The values of the Fourier coefficients of each semantic element are added to find the correlations between the given context words and each semantic element. The

semantic center $\mathbf{G}^{+}(s_\ell)$ of the sequence s_ℓ is then computed.

Step-3: Selection of the semantic elements which satisfy the criterion.

If the absolute value of the sum obtained in Step-2 for each individual semantic element is greater than a given threshold ε_s, those semantic elements are selected to form the semantic subspace.

Step-4: Selection of the positive or negative area of each of the selected semantic elements.

Each semantic element is divided into positive and negative areas, and the area which corresponds to the plus or minus sign of the value of the sum is selected. This area of the selected semantic element is employed to form the semantic subspace $P_\varepsilon(s_\ell)\mathcal{I}$.

5.2 Function-2: Selection of the Data Item Closest to the Keyword

The data item with the equivalent or closest meaning to the given keyword p is selected from the retrieval candidate data items of the given semantic subspace $P_\varepsilon(s_\ell)\mathcal{I}$.

When a keyword and a semantic subspace are given, the data items with the equivalent or closest meaning to the keyword in the given semantic subspace are selected from the retrieval candidate data item set $\mathcal{W}$.

In this system, the computation for obtaining the data item with the closest meaning to the keyword is performed very fast. This is because that the Fourier coefficients of

each data item can be computed and be sorted in the ascending order for each semantic element beforehand. This means that the data item with the closest meaning to the keyword is physically located at the closest position to that of the keyword in the specified semantic subspace. Therefore, the data item with the closest meaning to the keyword can be found very quickly without comparing the distances between the keyword and all the other data items.

This function is implemented by the following procedure. The correlations between the given keyword and each semantic element of the given semantic subspace are computed. That is, the keyword is mapped to the given semantic subspace.

Step-1: Mapping a keyword to the semantic subspace.

The Fourier expansion is computed for each element of the source vector of the given keyword. The Fourier coefficients are computed between each element of the vector and each semantic element involved in the given semantic subspace. Each coefficient is checked for whether or not it is in the semantic subspace, according to whether it is in the positive or negative area. If it is in the semantic subspace, the coefficient is added to the summation for the corresponding semantic element. Otherwise, the coefficient is ignored.

Step-2: Fourier expansion of the retrieval candidate data items.

This step is performed to map the candidate data items to the semantic subspace. The Fourier expansion is computed for the source vector of each candidate data

item, and the coefficients for each semantic element in the given semantic subspace are obtained.

Step-3: Computation of the distances between the keyword and retrieval candidate data items using the defined metric.

This step is performed to compute the distance between the keyword "p" to each retrieval candidate data item "w" by the metric defined in Subsection 3.2.5. As the result, the retrieval candidate data item with the closest meaning to the given keyword is selected from the specified candidate words in $\mathcal{W}$ in the given semantic subspace. This corresponds to finding the data item with the closest meaning to the given keyword from $\mathcal{W}$ in the given context. The system selects the data item which satisfies the following:

$$\min_{w \in \mathcal{W}} \rho\left(p, w; s_\ell\right).$$

5.3 Function-3: Selection of the Data Item From the Image Space

This function finds the data item with the closest meaning to the given keyword p from the data item set $\mathcal{W}$ in the whole semantic space, that is, the image space $\mathcal{I}$. This function is used to compute the meanings of the data items in context free cases.

When a keyword is given to the system, the data item with the closest meaning to the keyword in the image space is selected. This is a special case of Function-2 in which the semantic projection $P_\varepsilon(s_\ell)$ is the identity operator I.

6 Experiment

6.1 Recognition of Contexts

We performed five experiments (Experiment 1-1 to 1-5) in order to clarify the ability of our system to recognize contexts. These experiments were set up so as to evaluate the ability of our system to recognize "homonyms."

In these experiments, the first and the second retrieval candidate data items correspond to the same keyword. Keywords with two typical and different meanings are used. The first and the second retrieval candidate data items correspond to the first and the second meanings, respectively. Each keyword is used as a word which has two meanings, that is, the keyword is used as an ambiguous word.

The features of the keyword are the composite features created by composing the features of both the first and the second retrieval candidate data items. The first context is composed of the words which explain the first meaning of the keyword in the vocabulary entry of the dictionary. In this context, the keyword should be semantically close to the first retrieval candidate data item. The second context is composed of the words which explain the second meaning of the keyword in the vocabulary entry of the dictionary. In this context, the keyword should be semantically close to the second retrieval candidate data item.

The features of the first retrieval candidate data item are those of the vocabulary entry which explain the first retrieval candidate data item itself. These features corre-

spond to a row of the vocabulary entry of the first retrieval candidate data item in the data matrix A. The features of the second retrieval candidate data item are those of the vocabulary entry which explain the second retrieval candidate data item itself. These features correspond to a row of the vocabulary entry of the second retrieval candidate data item in the data matrix A.

Table 1 shows the keywords, the retrieval candidate data items and the given contexts. Table 2 shows the results of the experiments. These results show that our system has the ability to perform the dynamic retrieval necessary for precisely recognizing the different meanings of a single word according to the context. The results of the five experiments show that our system computes the meanings of data items reasonably.

In these experimental results, when the first context was given, the keyword was recognized as the word with the first meaning. Similarly, when the second context was given, the keyword was recognized as the second meaning. These results have evaluated the ability of our system in the recognition of contexts.

6.2 Experiments in a Multidatabase Environment

We performed several experiments at the attribute level by using two different databases which include bibliographical information. The objective of these experiments was to evaluate the effectiveness and applicability of our system to an actual multidatabase environment. In these different databases, several attributes with the same meaning are represented by different names. The system must be able to recognize "synonyms."

In the following experiments, it was assumed that the attribute names were not recognized by a user. The user issues a keyword and the context which gives the image of the keyword. The local keywords which correspond to the given keyword in each database are then extracted according to the given context by our system.

As an example, we used several attribute names as the retrieval candidate data items in two bibliographical databases named JMARC(JAPAN MARC; NATIONAL DIET LIBRARY) [5] and LCMARC (Library of Congress MARC U.S.Library) [6]. The JMARC database includes the bibliographical information of books published in Japan. The LCMARC database includes the same kind of information published in the United States of America. The metadatabase for these databases was created in the following way.

Step-1: Creation of the image space.

As described in Section 4, the image space is created by using the 2000 basic words. The image space is created independently of the contents of the individual databases. This space is used to map each attribute name with the database identifier to the image space itself.

Step-2: Definition of the additional explanatory words.

The explanatory words which are used to explain the attribute names are defined by using the 850 features which are used to define the columns of the data matrix A. If there are explanatory words which are not included in the 2000 vocabulary

entries, then those explanatory words are defined using the 850 features as additional explanatory words. As these additional words are not included in the 2000 words, these words are not used to create the image space. The additional explanatory words are dependent on the contents of the individual databases in the multidatabase environment. Each additional explanatory word is mapped to the image space by Fourier expansion.

In the experiments, as additional explanatory words, 40 words were defined. These additional explanatory words are used in the definitions of the attribute names in Step-3.

Step-3: Mapping attribute names to the image space.

In each database, each attribute name is explained using the basic and additional explanatory words. Each attribute name is mapped to the image space by applying Fourier expansion to each definition of those attribute names which are represented by the 850 features.

The attribute names which are used in the two different databases(JMARC and LCMARC) were mapped to the image space.

Step-4: Specification of a keyword, context words and a database identifier.

A keyword, object database identifiers and a context which fixes the semantic subspace are given. The context is given as a sequence consisting of basic words (vocabulary entries) and additional explanatory words.

The keyword is mapped to the image space by applying Fourier expansion to the definition of the keyword itself. In the semantic subspace, the closest data item(attribute name) to the keyword is then selected for each object database identifier.

The distances between a keyword and data items are computed in the semantic subspace corresponding to the given context, and the data item with the closest meaning to the keyword is selected for each object database identifier.

Table 3 shows the list of keywords and their explanations which are used as features. Tables 4 and 5 show the list of retrieval candidate data items, their explanatory words and the explanations used as features. In each experiment, database identifiers, a keyword, and context words were given, and the experimental results were obtained, as shown in Tables 6 and 7.

1. Experiment 2-1.

   In this experiment, the keyword "time" was issued to the system, and two different contexts were given. The first context fixed the selected semantic subspace consisting of the orthonormal bases of $\{\mathbf{q}_2, \mathbf{q}_{37}, \mathbf{q}_{51}, \mathbf{q}_{23}, \mathbf{q}_{97}, \cdots\}$. In this semantic subspace, in the case of the JMARC database, the three attribute names with the three shortest distances to the keyword "time" were extracted, and the attribute name "pubdate(published date)" was the selected data item. This means that the attribute name "pubdate" has the closest meaning to the keyword "time" in

JMARC in the first context. The attribute names "yr(year)" and "publisher" have the second and third closest meanings, respectively, in this context.

In the case of the LCMARC database with the same first context, the three attribute names with the three shortest distances to the keyword "time" were extracted. The attribute name "yr(year)" was the selected data item. This means that the attribute name "yr(year)" has the closest meaning to the keyword "time" in LCMARC in the given context.

The second context fixed the selected semantic subspace consisting of the orthonormal bases of $\{\mathbf{q}_2, \mathbf{q}_{37}, \mathbf{q}_{23}, \mathbf{q}_{97}, \mathbf{q}_{116}, \cdots\}$. In this semantic subspace, in the case of the JMARC database, the three attribute names with the three shortest distances to the keyword "time" were extracted, and the attribute name "yr(year)" was the selected data item. This means that the attribute name "yr(year)" has the closest meaning to the keyword "time" in JMARC in the second context.

In the case of the LCMARC database with the second context, the attribute names were extracted in the same way, and the attribute name "yr(year)" was the selected data item. This means that the attribute name "yr(year)" has the closest meaning to the keyword "time" in LCMARC in the second context. In LCMARC, the attribute name "yr(year)" is only the word related to the keyword "time" because the same attribute name "yr(year)" was selected in both the first

and second contexts.

This experment indicated that our system could dynamically extract the appropriate data item which had the closest meaning to the given keyword according to the given context. This dynamism was applied to the semantic interoperability of the experimental multidatabase environment, and the applicability of our system was evaluated.

2. Experiment 2-2.

   In this experiment, the keyword "write" was issued to the system, and a simple context was given. The context fixed the selected semantic subspace consisting of the orthonormal bases of $\{\mathbf{q}_9, \mathbf{q}_{13}, \mathbf{q}_{29}, \mathbf{q}_{19}, \mathbf{q}_{32}, \cdots\}$. In this semantic subspace, in the case of the JMARC database, the attribute name "author" was selected as the attribute name with the closest meaning to the keyword.

   In the case of the LCMARC database with this context, the attribute name "a(author)" was extracted in the same way.

3. Experiment 2-3.

   In this experiment, the keyword "pay" and the context word "cheap" were issued, and the semantic subspace consisting of the orthonormal bases of $\{\mathbf{q}_7, \mathbf{q}_{24}, \mathbf{q}_{60}, \mathbf{q}_{37}, \mathbf{q}_{17}, \cdots\}$ was selected. In both the JMARC and LCMARC databases, the attribute name "price" was extracted as the data item with the closest meaning to the keyword.

4. Experiment 2-4.

   In this experiment, the word "title" was issued as the keyword and the context word. The semantic subspace consisting of the orthonormal bases of $\{\mathbf{q}_{415}, \mathbf{q}_{446}, \mathbf{q}_{462}, \mathbf{q}_{431}, \mathbf{q}_{378}, \cdots\}$ is selected. In this subspace, the attribute name "title" was selected from the JMARC database, and the attribute name "t" was selected from the LCMARC database.

5. Experiment-2-5:

   In this experiment, the word "word" was issued as the keyword, and the context words "grammar" and "speech" were given. The semantic subspace consisting of the orthonormal bases of $\{\mathbf{q}_9, \mathbf{q}_7, \mathbf{q}_2, \mathbf{q}_{46}, \mathbf{q}_{76}, \cdots\}$ was selected. In this subspace, the attribute name "lg(text language)" was selected from the JMARC database, and the attribute name "lg(language)" was selected from the LCMARC database.

In our system, as shown by these experiments, the same concept with different representations among the different databases (for example, "author" in JMARC and "a(author)" in LCMARC) can be recognized according to the context, and the appropriate attribute names are extracted for each database. It is unnecessary for the users to know the local data representations in the local databases. The keyword and the context which are defined by the basic words and additional explanatory words can be used to issue queries. This fact reduces the overhead for issuing queries in a multidatabase environment.

This experimental study has clarified the applicability of our system to multidatabases for supporting semantic interoperability.

7 Conclusion

In this paper, we have proposed a fundamental framework for realizing semantic interoperability at the level of the semantic relationships between data items. Furthermore, as the implementation system of this framework, we have presented a metadatabase system for supporting semantic interoperability in a multidatabase environment. In this system, a mathematical model of meaning is used to realize semantic interoperability among the data items in multidatabases. This model is used as a basic computational model to extract the semantically equivalent data items which are included in different databases.

In this paper, we have assumed that a metadatabase has already been provided by using metainformation concerning each database to create the image space. In the future work, one of the most important issues is how to obtain the metainformation from various databases. We have classified the metainformation into three levels: the overview, attribute, and attribute-value levels. As the amount of metainformation in the attribute-value level will be very large, an automatic and dynamic acquisition mechanism for this information is required to realize an efficient metadatabase system.

We are currently extending our metadatabase system in order to apply it to more complicated semantic interoperability at the query transformation level. Furthermore,

we will also extend the framework for realizing more intelligent database environments. We have implemented a parallel processing system SMASH designed for advanced database applications [12, 13]. As the future work, the proposed metadatabase system will be implemented on SMASH.

References

[1] Batini, C.,Lenzelini, M. and Nabathe, S.B.(1986), "A comparative analysis of methodologies for database schema integration," *ACM Comp. Surveys,* Vol. 18, pp.323-364.

[2] Bright, M.W., Hurson, A.R., and Pakzad, S.H.(1992), "A Taxonomy and Current Issues in Multidatabase System," *IEEE Computer, Vol.25, No.3*, pp.50-59.

[3] Fang, D., Hammer, J., Mcleod, D.(1991), "The identification and resolution of semantic heterogeneity in multidatabase systems," *Proc. 1st IEEE Int. Workshop on Interoperability in Multidatabase Systems,* pp. 136-143.

[4] Gallant, S.I.(1991), "A practical approach for presenting context and for performing word sense disambiguation using neural networks," *Neural Computation,* 3, pp.293-309.

[5] JMARC : Japan MARC; *National Diet Library*, Japan.

[6] LCMARC: Library Congress Machine Readable Catalog, Library Congress, U.S.A.

[7] Larson, J.A., Navathe, S.B., Elmasri, R.(1989), "A theory of attribute equivalence in database with application to schema integration," *IEEE Transaction on Software Engineering* , Vol.15, No.4, pp.449-463.

[8] Litwin, W., Mark, L., and Roussopoulos, N.(1990), "Interoperability of Multiple Autonomous Databases," *ACM Comp. Surveys,* Vol.22, No.3, pp.267-293.

[9] *"Longman Dictionary of Contemporary English,"* Longman, (1987).

[10] Kitagawa, T. and Kiyoki, Y.(1993), "The mathematical model of meaning and its application to multidatabase systems," *Proc. 3rd IEEE Int. Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems,* pp.130-135.

[11] Kiyoki, Y. and Kitagawa, T.(1993), "A metadatabase system supporting interoperability in multidatabases,", *Proc. 1993 European-Japanese Seminar on Information Modeling and Knowledge Bases,* pp.484-495.

[12] Kiyoki, Y., Kurosawa, T., Kato, K. and Masuda, T.(1991), "The software architecture of a parallel processing system for advanced database applications," *Proc. 7th IEEE Int. Conference on Data Engineering,* pp. 220-229.

[13] Kiyoki, Y. and Namiuchi, M.(1991), "A stream-oriented parallel processing strategy for databases and knowledge bases," *Information Modelling and Knowledge Bases (IOS Press),* Vol. III, pp.316-332.

[14] Kolodner, J.L.(1984), "Retrieval and organizational strategies in conceptual memory: a computer model," *Lawrence Erlbaum Associates.*

[15] Ogden, C.K.(1940), "The General Basic English Dictionary," *Evans Brothers Limited.*

[16] Pu, C.(1990), "Semantic based integration library: A proposal for cooperative research for semantic interoperability," *Proc. Workshop on Multidatabases and Semantic Interoperability,* pp.6-9.

[17] Sheth, A. and Larson, J.A.(1990), "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Computing Surveys,* Vol.22, No.3, pp.183-236.

[18] Sheth, A. and Kashyap, V.(1992), *"So far (schematically) yet so near (semantically),"* Proc. IFIP TC2/WG2.6 Conf. on Semantics of Interoperable Database Systems, pp.1-30.

[19] Shimizu, H., Kiyoki, Y., Sekijima, A. and Kamibayashi, N.(1991), "A Decision Making Support System for Selecting Appropriate Online Databases,", *Proc. 1st IEEE Int. Workshop on Interoperability in Multidatabase Systems,* pp.322-329.

[20] Yu, C., Sun, W., Dao, S., Keirsey, D.(1990), "Determining relationships among attributes for interoperability of multi-database systems," *Proc. Workshop on Multidatabases and Semantic Interoperability,* pp.10-15.